CHAPTER 2: REVIEW OF THE CAHSEE AGAINST STANDARDS FOR TEST DEVELOPMENT

Introduction

The first question asked in AB 1609 is whether development of the CAHSEE meets the standards for a test of this type. Analyses of the appropriateness and quality of the exams developed to date have been a major focus of ongoing evaluation efforts (e.g., Wise et al., 2002).

Standards for test development have been prepared by joint committees of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (Standards for educational and psychological testing referred to here simply as the Standards). The most recent edition of these standards was published in 1999 (AERA, APA, NCME, 1999). These Standards are widely accepted as the most comprehensive and authoritative statement of standards for educational tests.

In this chapter, the relevant standards for educational and psychological tests are listed and findings from the ongoing evaluation are used to determine the extent to which the CAHSEE meets these standards. The chapter is organized into several sections that cover standards for different aspects of test development and use, beginning with standards for validity. The chapter concludes with a summary of overall findings and a discussion of a few overarching issues.

Note that the *Standards* cover a wide range of tests and testing situations. Not all standards are relevant to the CAHSEE. Many cover issues unique to areas such as employment testing or the use of tests in making psychological diagnoses. In the interests of brevity and clarity, we have not included the irrelevant standards nor discussed why they do not apply to the CAHSEE. So, in the following text where standard numbers are not consecutive, the skipped standards were not considered relevant. In a few cases, standards of possible relevance are listed and the conclusion that they are not relevant to the primary use of the CAHSEE as a requirement for high school graduation is discussed.

In addition to information from prior evaluation reports, we reviewed technical documentation provided by the test development contractors. These documents include Yoon and Williams (2002), Smith, Suh, Yoon, and Williams (2002), and Educational Testing Service (2003).

Test Construction, Evaluation, and Documentation

The first part of the *Standards* covers standards for the development and documentation of tests. Validity and reliability are the two most central issues. These are covered first, followed by discussion of standards for development, administration, and reporting.

Standards for Validity

As described in the *Standards*, "Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of the test." For the CAHSEE, test scores are interpreted as an indicator of whether students have or have not mastered the content set by policy as a graduation requirement. The standards for validity that the CAHSEE must meet are shown and discussed in this section.

Validity may be established in a number of ways, depending on the nature and use of the test. Since the CAHSEE is not explicitly used to predict future outcomes (as are employment tests), the validity of the CAHSEE as a graduation requirement is established through expert judgment of the extent to which the CAHSEE scores accurately reflect mastery of the content established by the State Board of Education (SBE) as required for graduation. Other forms of validation, such as predictive validity studies or analyses of relationships between the CAHSEE scores and other measures of the same construct are not appropriate. Standards for these other forms of validation are omitted from the present discussion.

Standard 1.1: A rationale should be presented for each recommended interpretation and use of test scores, together with a comprehensive summary of the evidence and theory bearing on the intended use or interpretation.

The rationale for the use of the CAHSEE as a high school graduation requirement was initially specified in the legislation establishing the exam. (See California Education Code, Chapter 8, Section 60850.) The content covered by the CAHSEE was adopted by the SBE at its October 2000 meeting, following recommendations from an independent panel of experts. Criteria for demonstrating adequate proficiency in the two sections of the exam (English-Language Arts and Mathematics) were also adopted by SBE (June 2001) based on recommendations from another independent panel. Both actions were public and well documented in the minutes of the SBE. The California Department of Education (CDE) has subsequently published descriptions of the CAHSEE and its use on the CDE website (http://www.cde.ca.gov/statetests/cahsee/). **Standard 1.1 is fully met.**

Standard 1.2: The test developer should set forth clearly how test scores are intended to be interpreted and used. The population(s) for which a test is appropriate should be clearly delimited, and the construct that the test is intended to assess should be clearly described.

The intended interpretation and use of the CAHSEE scores in determining whether students meet diploma requirements is specified in EC 60854. The CAHSEE scores are intended to represent mastery of specified content standards in English-language arts and mathematics. These content standards are laid out in blueprints for the exam adopted by SBE (http://www.cde.ca.gov/statetests/cahsee/admin/blueprints/langarts.pdf; http://www.cde.ca.gov/statetests/cahsee/admin/blueprints/math.pdf). In the initial legislation, and in subsequent documents published by CDE, the target populations are clearly specified as successive classes of California high school students beginning with the Class of 2004. **Standard 1.2 is fully met.**

Standard 1.6: When the validation rests in part on the appropriateness of test content, the procedures followed in specifying and generating test content should be described and justified in reference to the construct the test is intended to measure or the domain it is intended to represent. If the definition of the content sampled incorporates criteria such as importance, frequency, or criticality, these criteria should be clearly explained and justified.

Each test question is targeted to a particular standard in the content domain established by SBE. The relevance of the question to the targeted content standard is checked by the item writer, content editors, and independent review panels as described in technical documentation provided by the test development contractors (Yoon & Williams, 2002; Smith et al., 2002; Educational Testing Service, 2003). In conducting the independent evaluation of the CAHSEE, HumRRO has twice convened additional panels to check the procedures used by the developer for generating appropriate test questions for each of the targeted content standards (Wise et al., 2000; Wise et al., 2002b). Results of these independent checks corroborated the validity of the process used by the developers. **Standard 1.6 is fully met.**

Standard 1.7: When a validation rests in part on the opinions or decisions of expert judges, observers, or raters, procedures for selecting such experts and for eliciting judgments or ratings should be fully described. The qualifications, and experience, of the judges should be presented. The description of procedures should include any training and instructions provided, should indicate whether participants reached their decisions independently, and should report the level of agreement reached. If participants interacted with one another or exchanged information, the procedures through which they may have influenced one another should be set forth.

The processes used by the developers to review each test question, for conducting expert reviews of test content and the composition of the panels used with these procedures are described fully in their technical documentation. The selection and characteristics of the panels used by the independent evaluator in confirming the appropriateness of the CAHSEE test questions for the targeted content also are fully documented in the evaluation reports. **Standard 1.7 is fully met.**

Standard 1.10: When interpretation of performance on specific items, or small subsets of items, is suggested, the rationale and relevant evidence in support of such interpretation should be provided. When interpretation of individual item responses is likely but is not recommended by the developer, the user should be warned against making such interpretations.

The determination of whether a student does or does not pass each part of the CAHSEE is based on all of the ELA or mathematics items, not on specific items or on small sets of items. Thus **Standard 1.10 does not apply to the CAHSEE** in its use as a graduation requirement. Note, however, that the CAHSEE score reports do include information on performance on items in each content strand. In many cases, the number of items covering a given strand is limited. CDE may wish to consider further review of whether appropriate caution is given for the interpretation of these subscores.

Standard 1.12: When interpretation of subscores, score differences, or profiles is suggested, the rationale and relevant evidence in support of such interpretation should be provided. Where composite scores are developed, the basis and rationale for arriving at the composites should be given.

Separate scores for English-language arts and mathematics are used in determining eligibility for high school graduation. Each of these scores is a composite of scores on questions assessing the individual standards. The method and rationale for computing the composites follows the test blueprints adopted by the SBE. **Standard 1.12 is fully met** with respect to use of the CAHSEE scores as a high school graduation requirement.

Standard 1.22. When it is clearly stated or implied that a recommended test use will result in a specific outcome, the basis for expecting that outcome should be presented, together with relevant evidence.

The use of the CAHSEE scores in determining eligibility for a diploma is set by policy and not explicitly tied to specific outcomes. The implied outcome is that students receiving a high school diploma will possess core skills essential for success in school, work, or other activities following high school. The basis for expecting this outcome is the judgment of the HSEE Standards Panel[1] and the SBE in establishing the content and performance standards measured by the CAHSEE. Empirical checks on these judgments cannot be conducted until the CAHSEE has been in place for some period of time, thus we conclude that **Standard 1.22 is not relevant at this time.**

Standard 1.23: When a test use or score interpretation is recommended on the grounds that testing or the testing program per se will result in some indirect benefit in addition to the utility of information from the test scores themselves, the rationale for anticipating the indirect benefit should be made explicit. Logical or theoretical arguments and empirical evidence for the indirect benefit should be provided. Due weight should be given to any contradictory findings in the scientific literature, including findings suggesting important indirect outcomes other than those predicted.

A clearly implied benefit from the imposition of the CAHSEE graduation requirement is that instruction in knowledge and skills deemed essential will be significantly improved. The validity of this assumption was a target of the current investigation. Empirical support for this assumption is provided in Chapter 3 of this report. **Standard 1.23 is fully met.**

Standard 1.24: When unintended consequences result from test use, an attempt should be made to investigate whether such consequences arise from the test's sensitivity to characteristics other than those it is intended to assess or to the test's failure fully to represent the intended construct.

[1] Following provisions in the legislation, a panel of teachers, principals, school board members, parents, and the general public was appointed by the Superintendent and approved by the Board. The HSEE Standards Panel's primary responsibility is to ensure that the exam is aligned with the Board's rigorous content standards for ELA and mathematics (EC60850b). The Panel also considers and makes recommendations on a range of test development and administration issues such as frequency of testing, accommodations for students with disabilities, and determination of passing levels.

A major goal of the ongoing independent evaluation of the CAHSEE is to identify any unintended consequences resulting from the use of the CAHSEE as a graduation requirement. To date, no unintended consequences have been found, although it should be noted that no student has yet been denied a diploma. In addition, the development contractor, other experts selected by SBE and CDE, and the independent evaluators have reviewed test questions to ensure that they do not require extraneous knowledge or skill and do represent content standards fully. Thus, even if unintended consequences should arise, it is highly unlikely that they will be associated with either inappropriate or incomplete measurement. **Standard 1.24 is fully met at this time.**

Standards for Test Reliability

Whereas validity concerns the extent to which tests measure content appropriate to the interpretations made of the test scores, reliability concerns the accuracy with which such content is measured. Accuracy can be described in a number of ways. Traditional reliability coefficients estimate the degree to which (percent of) variation in test scores is repeatable across independent assessments. Standard errors of measurement assess the extent of variation in test scores for a given individual that would likely result from independent administrations of the test. In the present context, the CAHSEE scores are used to classify students as either passing the graduation standards or failing to meet these standards. In this case, the accuracy of such classifications overall, and for students in different score ranges, is a primary reliability issue.

Standard 2.1: For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant reliabilities and standard errors of measurement or test information functions should be reported.

Reliability estimates and standard error of measurement for each test form are included in the technical documentation provided by the test development contractor. In addition, the evaluation reports include extensive analyses of classification accuracy. **Standard 2.1 is fully met with respect to the use of the CAHSEE scores as a graduation requirement.**

It should be noted that subscores for each content area are reported for diagnostic use. The reliability of these scores can vary across different forms of the test. Reliability estimates for these scores are included in the technical documentation (Educational Testing Service, 2003, p. 105), although the analysis of subscore reliability is not extensive.

Standard 2.2: The standard error of measurement, both overall and conditional (if relevant), should be reported both in raw score or original scale units and in units of each derived score recommended for use in test interpretation.

Reliability of the CAHSEE scores has primarily been reported with respect to the derived scale used for reporting. Analyses by the evaluator have indicated error bands with respect to the percent of items answered correctly as well as with respect to this reporting scale. **Standard 2.2 is adequately met.**

Standard 2.3: When test interpretation emphasizes differences between two observed scores of an individual or two averages of a group, reliability data, including standard errors, should be provided for such differences.

Interpretation of score differences is not relevant to the use of the CAHSEE as a high school graduation requirement, thus **Standard 2.3 is not relevant to the primary use of the CAHSEE**. Nonetheless, students repeating the CAHSEE receive multiple scores and interpretation of differences in these scores by the students and their parents and teachers is likely. Evaluation results suggest some issues with the accuracy of individual change scores for some parts of the score scale (Wise et al., 2002). Cautions based on these findings are being considered.

The CAHSEE scores also are being used for school accountability purposes. Comparisons across schools and districts are inevitable when accountability results are presented. Analyses of the accuracy of such comparisons have been conducted by CDE but are outside the scope of the current investigation.

Standard 2.4: Each method of quantifying the precision or consistency of scores should be described clearly and expressed in terms of statistics appropriate to the method. The sampling procedures used to select examinees for reliability analyses and descriptive statistics on these samples should be reported.

The procedures and samples used by the test development contractor in estimating reliability coefficients and standard error of measurement for each test form are described completely in their technical documentation. The procedures and samples used by the evaluator in analyses of classification accuracy are similarly well described. **Standard 2.4 is fully met.**

Standard 2.7: When subsets of items within a test are dictated by the test specifications and can be presumed to measure partially independent traits or abilities, reliability estimation procedures should recognize the multifactor character of the instrument.

The CAHSEE covers content standards organized into a number of discrete areas or strands. Scores for separate strands are reported for diagnostic use. The psychometric model used for the overall scores assumes that results for each standard are indicators of performance on a single underlying dimension of achievement. Thus **Standard 2.7 is not relevant to the use of the CAHSEE in determining eligibility for a diploma.**

Standard 2.8: Test users should be informed about the degree to which rate of work may affect examinee performance.

Standard 2.9: When a test is designed to reflect rate of work, reliability should be estimated by the alternate-form or test-retest approach, using separately timed administrations.

Students are given essentially unlimited time to complete each portion of the exam. Consequently rate of work is not part of the construct being measured and **Standards 2.8 and 2.9 do not apply to the CAHSEE**.

Standard 2.10: When subjective judgment enters into test scoring, evidence should be provided on both inter-rater consistency in scoring and within-examinee consistency over repeated measurements. A clear distinction should be made among reliability data based on (a) independent panels of raters scoring the same performances or products, (b) a single panel scoring successive performances or new products, and (c) independent panels scoring successive performances or new products.

Responses to the essay questions in the ELA exam are rated by scorers following specified rubrics for judging these responses. Each response is independently rated by two different scorers. This provides a basis for establishing the consistency of scoring judgments. Analyses of inter-rater consistency are reported in the test development contractors' technical documentation and have also been analyzed in the evaluation reports. **Standard 2.10 is fully met.**

Standard 2.11: If there are generally accepted theoretical or empirical reasons for expecting that reliability coefficients, standard errors of measurement, or test information functions will differ substantially for various subpopulations, publishers should provide reliability data as soon as feasible for each major population for which the test is recommended.

Information on the CAHSEE score accuracy is based on Item Response Theory (IRT) models in which performance on the test questions and the exam as a whole has the same functional relationship to the underlying trait being measured for all groups. Individual test questions are screened for differential item functioning (DIF) using checks to see that performance on the questions is not differentially related to membership in racial/ethnic groups or to gender. **Standard 2.11 is adequately met**, although CDE may wish to consider additional analyses of reliabilities for targeted subgroups.

Standard 2.14: Conditional standard errors of measurement should be reported at several score levels if constancy cannot be assumed. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score.

Standard 2.15: When a test or combination of measures is used to make categorical decisions, estimates should be provided of the percentage of examinees who would be classified in the same way on two applications of the procedures, using the same form or alternate forms of the instrument.

Technical documentation of each test form includes estimates of standard errors of measurement for different score levels. Classification accuracy overall, and for students in different score ranges, has been estimated by the evaluation contractor. **Standards 2.14 and 2.15 have been fully met.**

Standard 2.18: When significant variations are permitted in test administration procedures, separate reliability analyses should be provided for scores produced under each variation if adequate sample sizes are available.

Some variations in test administration procedures are allowed to accommodate students with special needs whose Individualized Education Plans (IEPs) or Section 504 plans, specify the need for such accommodations. In all cases, the accommodations have been judged to not alter the construct being assessed. The number of students allowed specific accommodations has been too small to permit separate estimates of reliability to be computed accurately. Thus, **Standard 2.18 is not relevant at this time**. As additional data are collected, CDE may wish to investigate further the reliability of scores for students requiring specific accommodations.

Standard 2.19: When average test scores for groups are used in program evaluations, the groups tested should generally be regarded as a sample from a larger population, even if all examinees available at the time of measurement are tested. In such cases the standard error of the group mean should be reported, as it reflects variability due to sampling of examinees as well as variability due to measurement error.

Standard 2.19 does not apply to the use of the CAHSEE as a high school graduation requirement. Treatment of existing students as a sample of a larger population is a consideration in the use of the CAHSEE scores for accountability.

Test Development and Revision

Standard 3.1: Tests and testing programs should be developed on a sound scientific basis. Test developers and publishers should compile and document adequate evidence bearing on test development.

Test development procedures have been reviewed extensively by internal and outside experts and are fully documented in technical reports provided by the development contractor. **Standard 3.1 is fully met.**

Standard 3.2: The purpose(s) of the test, definition of the domain, and the test specifications should be stated clearly so that judgments can be made about the appropriateness of the defined domain for the stated purpose(s) of the test and about the relation of items to the dimensions of the domain they are intended to represent.

The purpose and general domain of the test are clearly specified in the enabling legislation. Specific descriptions for the domain covered by the test and specifications for coverage of each area are provided in test blueprints. These documents are publicly available in printed form through the CDE website. **Standard 3.2 is fully met.**

Standard 3.3: The test specifications should be documented, along with their rationale and the process by which they were developed. The test specifications should define the content of the test, the proposed number of items, the item formats, the desired psychometric properties of the items, and the item and section

arrangement. They should also specify the amount of time for testing, directions to the test takers, procedures to be used for test administration and scoring, and other relevant information.

A detailed and public process was used in specifying the content of the CAHSEE. As required in the enabling statutes, an advisory committee was formed, held extensive public hearings, and recommended content standards for the CAHSEE to the SBE. The State Board made final decisions on the standards to be tested and approved blueprints specifying the number and types of questions for each standard. The specifications and blueprints have been published on the Department's website and widely distributed. Documents describing test administration procedures are also posted on the Department's website and have been distributed to testing coordinators for each district. These documents are reviewed during training for test administrators. **Standard 3.3 is fully met.**

Standard 3.4: The procedures used to interpret test scores, and, when appropriate, the normative or standardization samples or the criterion used should be documented.

The primary use of the test scores is to determine whether students have or have not achieved sufficient mastery of the targeted content standards to be granted a diploma. The criterion for passing each test, specified in terms of the number and percent of items on the original form answered correctly, was adopted by the State Board of Education and is recorded in the SBE minutes as well as in several documents published by the Department and its test development contractor. Score reports clearly indicate whether students did or did not meet the passing criteria. **Standard 3.4 is fully met.**

Note that the test is not intended to be used to compare a student's performance to that of other students. Norms showing the percentage of students in some reference population at or above each score level have not been published.

As students who do not pass on their first try subsequently retake one or both parts of the exam, another interpretive use arises. Students, parents, and teachers will seek to interpret differences between a student's original and subsequent scores on the underlying reporting scale. As noted by Wise et al. (2002), the impact of guessing with multiple choice questions may confound the interpretation of gain scores for students at very low score levels. The Department may wish to develop and distribute more detailed guidelines for interpreting gain scores.

Score reports also include information on the number of questions in each major content area (strand) that are answered correctly. These numbers can be compared to the total number of questions in each area. Interpretation of these numbers may be limited by differences in the relative difficulty of questions in different strands. Normative information on the subscores is provided in technical documentation for each test form. CDE may wish to develop more specific guidance for use and interpretation of subscores as indicators of a student's relative strengths and weaknesses.

Standard 3.5: When appropriate, relevant experts external to the testing program should review the test specifications. The purpose of the review, the processes by which the review is conducted, and the results of the review should be documented. The qualifications, relevant experience, and demographic characteristics of expert judges should also be documented.

The CAHSEE Panel convened to develop the test specifications as required in the enabling legislation included individuals with specific expertise in different aspects of instruction and testing. The composition and deliberations of this panel are well documented in the panel minutes and in the final report to SBE. In addition, the panel assembled technical committees with specific expertise in mathematics and in English-language arts. The recommendations of the technical committees and of other experts who participated in public hearings also are documented in minutes from the panel meetings. **Standard 3.5 is fully met.**

Standard 3.6: The type of items, the response formats, scoring procedures, and test administration procedures should be selected based on the purposes of the test, the domain to be measured, and the intended test takers. To the extent possible, test content should be chosen to ensure that intended inferences from test scores are equally valid for members of different groups of test takers. The test review process should include empirical analyses and, when appropriate, the use of expert judges to review items and response formats. The qualifications, relevant experiences, and demographic characteristics of expert judges should also be documented.

Essay questions are used to assess writing skills and multiple-choice questions are used to assess mastery of other content standards. These response formats are both common and appropriate for the target population of high school students.

The specific questions included in each test form are reviewed extensively by contractor staff, outside panels, CDE, and ultimately SBE. Reviews include consideration of whether the question is an appropriate measure of the targeted content standard. Specific review for bias and fairness is an integral part of this process. Procedures for conducting bias and fairness reviews developed by ETS are an industry standard. Review of item statistics for any differential functioning across examinee groups also follows industry standards developed by ETS. Both the content and statistical review procedures have been followed by the original test development contractor and now by ETS. Description of the procedures and the reviewers included at each stage are included in the contractor's technical documentation. **Standard 3.6 is fully met.**

Standard 3.7: The procedures used to develop, review, and try out items, and to select items from the item pool should be documented. If the items were classified into different categories or subtests according to the test specifications, the procedures used for the classification and the appropriateness and accuracy of the classification should be documented.

Item development and review procedures are included in the development contractor's technical documentation. Each question is subjected to a field test and statistical results from the field test are used in further screening potential questions before they are added to the

bank of available questions. The development contractor also provides documentation of the results from each field test. In addition to reviews conducted by CDE and its contractors, HumRRO's evaluation of the CAHSEE has included two independent reviews of the match of test questions to content standards (Wise et al., 2000; Wise et al., 2002). Results indicated that the contractor's development and review procedures were working appropriately. **Standard 3.7 is fully met.**

Standard 3.8: When item tryouts or field tests are conducted, the procedures used to select the sample(s) of test takers for item tryouts and the resulting characteristics of the sample(s) should be documented. When appropriate, the sample(s) should be as representative as possible of the population(s) for which the test is intended.

Standard 3.9: When a test developer evaluates the psychometric properties of items, the classical or item response theory (IRT) model used for evaluating the psychometric properties of items should be documented. The sample used for estimating item properties should be described and should be of adequate size and diversity for the procedure. The process by which items are selected and the data used for item selection, such as item difficulty, item discrimination, and/or item information, should also be documented. When IRT is used to estimate item parameters in test development, the item response model, estimation procedures, and evidence of model fit should be documented.

All questions used in operational forms of the CAHSEE are first included in a field test. With two exceptions, field test questions are embedded within operational test forms so that the field test samples and testing conditions are identical to operational conditions. The only exceptions are the initial field tests conducted in 2000 before any operational forms were assembled and a subsequent field test of essay questions. (Tryout versions of essay questions cannot be embedded into operational test forms because of time considerations.) In all cases, the contractor has documented the characteristics of the field test samples, and they have been judged to be appropriate by the independent evaluators.

Procedures used to review field test results include analysis of both classical item statistics and IRT parameter estimates. The procedures and their results are included in technical documentation provided by the contractor. **Standards 3.8 and 3.9 are fully met.**

Some of the field tests have been based on examinees retaking the CAHSEE for the second or subsequent times. While these examinees are a key part of the target population for the CAHSEE, their performance is typically lower than that of first-time test takers in general. As new classes of students begin to take the CAHSEE, the number of first-time test takers will increase dramatically. CDE may wish to restrict future field tests to first-time test takers. Test forms designed exclusively for retest situations might include additional equating questions rather than field test questions.

Standard 3.11: Test developers should document the extent to which the content domain of a test represents the defined domain and test specifications.

As noted above, test blueprints indicate intended coverage of each content standard and all test forms follow these blueprints. **Standard 3.11 is adequately met**. Educational researchers are developing measures of alignment between tests and content standards with more sensitive measures of the coverage of specific standards by test questions. One example is the depth of content dimension in Webb's (2002) model. CDE and the development contractors may wish to explore the appropriateness of such approaches for identifying particular content standards that are difficult to assess completely with an aim toward either revising the content descriptions or expanding item types to cover the content more fully.

Standard 3.13: When a test score is derived from the differential weighting of items, the test developer should document the rationale and process used to develop, review, and assign item weights. When the item weights are obtained based on expert judgment, the qualifications of the judges should be documented.

With the exception of the essay questions, the test questions are given equal weight. Specifications for the relative weight given to the essay questions were developed in consultation with the CAHSEE Panel and approved by SBE. The qualifications of the participants in this process are fully documented in Panel and SBE minutes. **Standard 3.13 is fully met.**

Standard 3.14: The criteria used for scoring test takers' performance on extended-response items should be documented. This documentation is especially important for performance assessments, such as scorable portfolios and essays, where the criteria may not be obvious to the user.

The test publisher documents general descriptions of score levels for the essay questions and the specific rubrics used with each individual question. The independent evaluators have reviewed the scoring procedures and have made minor suggestions for improvement. **Standard 3.l4 is fully met.**

Standard 3.19: The directions for test administration should be presented with sufficient clarity and emphasis so that it is possible for others to replicate adequately the administration conditions under which the data on reliability and validity, and, where appropriate, norms were obtained.

First, note that the validity of the CAHSEE scores for high school graduation has been established through expert judgment about test content that does not depend on actual administration of the test. Similarly, normative information is not relevant to the use of the CAHSEE as a high school graduation requirement. Nonetheless, test administration procedures have been documented in detail and training has been provided to testing coordinators. **Standard 3.19 is fully met.**

Standard 3.20: The instructions presented to test takers should contain sufficient detail so that test takers can respond to a task in the manner that the test developer intended. When appropriate, sample material, practice or sample questions, criteria for scoring, and a representative item identified with each major area in the test's classification or domain should be provided to the test takers prior to the

administration of the test or included in the testing material as part of the standard administration instructions.

The instructions to test takers have been reviewed by CDE and SBE staff, as well as by the test developer's technical experts. Sample questions were released before the first administration of the CAHSEE and additional questions are being released each year. The released questions are linked to the test content standards. One minor concern with the instructions to test takers was noted in the Year 3 report from the independent evaluation (Wise et al. June 2002). Scoring rubrics for some of the essay questions include evaluation of whether the response is appropriate for the intended audience, but the question posed to the test takers has not always indicated an audience for the student's response. The test developer is addressing this issue. With this one minor adjustment, **Standard 3.20 is fully met.**

Standard 3.21: If the test developer indicates that the conditions of administration are permitted to vary from one test taker or group to another, permissible variation in conditions for administration should be identified, and a rationale for permitting the different conditions should be documented.

The development contractor, CDE staff, and SBE have given careful consideration to allowable accommodations. These accommodations are consistent with common practice and constitute the only permissible variation in administration procedures. The rationale for these variations is included in the regulations for test accommodations. **Standard 3.21 is fully met.**

Standard 3.22: Procedures for scoring and, if relevant, scoring criteria should be presented by the test developer in sufficient detail and clarity to maximize the accuracy of scoring. Instructions for using rating scales or for deriving scores obtained by coding, scaling, or classifying constructed responses should be clear. This is especially critical if tests can be scored locally.

The ELA portion of the CAHSEE includes two essays that are scored by the test development contractor. Scoring criteria for each question are developed and thoroughly reviewed before test forms are printed. Scorer training is documented, and there are extensive quality-control checks on scoring accuracy both before and during operational scoring. All essays are scored by two independent scorers and, if significant disagreements are found, by one or two additional scorers. **Standard 3.22 is fully met.**

Standard 3.23: The process for selecting, training, and qualifying scorers should be documented by the test developer. The training materials, such as the scoring rubrics and examples of test takers' responses that illustrate the levels on the score scale, and the procedures for training scorers should result in a degree of agreement among scorers that allows for the scores to be interpreted as originally intended by the test developer. Scorer reliability and potential drift over time in raters' scoring standards should be evaluated and reported by the person(s) responsible for conducting the training session.

Technical documentation supplied by the test developer described the process for selecting, training, and monitoring scoring of the essays. Monitoring procedures include recalibration exercises at the beginning of each scoring session and periodic checks for scoring drift. Agreement statistics have been analyzed and reported both by the developer and the evaluator. **Standard 3.23 is fully met.**

Scales, Norms, and Score Comparability

Section 4 of the *Standards* covers the development of scales used for reporting test results along with the creation and documentation of information to support interpretations of these scores.

Standard 4.1: Test documents should provide test users with clear explanations of the meaning and intended interpretation of derived score scales, as well as their limitations.

The score scale used for reporting was designed to run from 250 to 450. The scale was adjusted so that the passing level would be at 350 and the point corresponding to chance responding on the multiple choice questions would be 300. Information on guessing levels is described in detail only in technical documentation, but the passing level is clearly communicated in all documents describing test results. Since the interpretation of scores with respect to the passing level is the primary use intended for these scores, **Standard 4.1 is adequately met.** More information on the guessing levels might be provided to users to avoid possible misinterpretation of scores and score gains below the chance level.

Standard 4.2: The construction of test scales used for reporting scores should be described clearly in test documentation.

Conversion tables showing how reported scale scores are derived from the raw score have been provided for each test form. The raw score is simply the number of correct responses for mathematics. For ELA, the raw score is a weighted sum of the number of correct responses to the multiple-choice questions and the scores on the two essay questions. **Standard 4.2 is fully met.**

Standard 4.3: If there is sound reason to believe that specific misinterpretations of a score scale are likely, test users should be explicitly forewarned.

The score scale used for reporting was developed to provide a constant interpretation of test scores across test forms that vary slightly in difficulty. The nature of this scale is explained in the score reports and tables for converting number correct scores from a given form onto the reporting scale are provided in technical documentation, along with estimates of error or measurement. Limitations on the interpretations of scores at the low end of the scale due to the impact of guessing have been reported by the evaluator. **Standard 4.3 is adequately met.**

Standard 4.9: When raw score or derived score scales are designed for criterion-referenced interpretation, including the classification of examinees into separate

categories, the rationale for recommended score interpretations should be clearly explained.

The reporting scale has been designed so that 350 is always the minimum passing score for each test. Mastery of the required content was initially also defined in terms of minimum percent correct scores (60 percent for ELA and 55 percent for mathematics), although this varies slightly across test forms. No further explanation is required and **Standard 4.9 is fully met**.

Standard 4.10: A clear rationale and supporting evidence should be provided for any claim that scores earned on different forms of a test may be used interchangeably. … The specific rationale and the evidence required will depend in part on the intended uses for which score equivalence is claimed.

Standard 4.11: When claims of form-to-form equivalence are based on equating procedures, detailed technical information should be provided on the method by which equating functions or other linkages were established and on the accuracy of equating functions.

Standard 4.13: In equating studies that employ an anchor test design, the characteristics of the anchor test and its similarity to the forms being equated should be presented, including both content specifications and empirically determined relationships among test scores. If anchor items are used, as in some IRT-based and classical equating studies, the representativeness and psychometric characteristics of anchor items should be presented.

Standard 4.17: Testing programs that attempt to maintain a common scale over time should conduct periodic checks of the stability of the scale on which scores are reported.

Each test form is built to the same test blueprint, specifying the required number of questions for each content standard. Technical documentation for each administration includes an extensive discussion and analysis of test form equating procedures and results. Equating procedures involve the use of a substantial number of anchor questions that cover each subscale of each of the tests. In all cases, equating results have supported the equivalence of the resulting scale scores. These results have been reviewed by the independent evaluator and by other outside technical experts. **Standards 4.10, 4.11, 4.13, and 4.17 are fully met.**

Standard 4.19: When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be clearly documented.

Standard 4.21: When cut scores defining pass-fail or proficiency categories are based on direct judgments about the adequacy of item or test performances or performance levels, the judgmental process should be designed so that judges can bring their knowledge and experience to bear in a reasonable way.

Procedures used to develop recommended cut scores have been documented in a technical report provided by the development contractor. Industry standard procedures for selecting panelists and for eliciting valid judgments from them were employed. Considerations by the SBE in making final decisions on the cut scores that define passing levels for each of the two content areas are documented in SBE meeting minutes. **Standards 4.19 and 4.21 are fully met.**

Test Administration, Scoring, and Reporting

Section 5 of the *Standards* covers additional issues in the administration and use of tests. Relevant standards are listed here, although there is considerable overlap with the standards for test development discussed above.

Standard 5.1: Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer, unless the situation of a test taker's disability dictates that an exception should be made.

Test administration manuals describing prescribed testing procedures have been developed by the contractor responsible for development, administration, and scoring of the CAHSEE. Training workshops are provided for testing coordinators during which the test administration manuals are reviewed in detail. **Standard 5.1 is fully met.**

Standard 5.2: Modifications or disruptions of standardized test administration procedures or scoring should be documented.

Standard 5.3: When formal procedures have been established for requesting and receiving accommodations, test takers should be informed of these procedures in advance of testing.

Test accommodations and procedures for requesting them have been established in regulations adopted by SBE. Letters were sent to parents of students in the Class of 2004 informing them of students' rights to these accommodations. **Standards 5.2 and 5.3 are fully met.**

Standard 5.4: The testing environment should furnish reasonable comfort with minimal distractions.

Guidance for testing environments is provided to local testing coordinators. Test administrations are monitored at a sample of sites. Insofar as can be determined, **Standard 5.4 is adequately met**.

Standard 5.5. Instructions to test takers should clearly indicate how to make responses. Instructions should also be given in the use of any equipment likely to be unfamiliar to test takers. Opportunity to practice responding should be given when equipment is involved, unless use of the equipment is being assessed.

No equipment, beyond a number 2 pencil, is required. Instructions for marking responses are provided. **Standard 5.5 is fully met**.

Standard 5.6: Reasonable efforts should be made to assure the integrity of test scores by eliminating opportunities for test takers to attain scores by fraudulent means.

Standard 5.7: Test users have the responsibility of protecting the security of test materials at all times.

Test administration procedures include very explicit instructions for protecting the security of test materials as well as for preventing or detecting various forms of cheating. **Standards 5.6 and 5.7 are fully met.**

Standard 5.8. Test scoring services should document the procedures that were followed to assure accuracy of scoring. The frequency of scoring errors should be monitored and reported to users of the service on reasonable request. Any systematic source of scoring errors should be corrected.

Standard 5.9: When test scoring involves human judgment, scoring rubrics should specify criteria for scoring. Adherence to established scoring criteria should be monitored and checked regularly. Monitoring procedures should be documented.

Procedures for monitoring the accuracy of scoring, particularly the scoring of student essays, are described in the technical documentation, along with analyses of resulting scoring accuracy. **Standards 5.8 and 5.9 are fully met.**

Standard 5.10: When test score information is released to students, parents, legal representatives, teachers, clients, or the media, those responsible for testing programs should provide appropriate interpretations. The interpretations should describe in simple language what the test covers, what scores mean, the precision of the scores, common misinterpretations of test scores, and how scores will be used.

Score reports provided to students and parents have been thoroughly reviewed for completeness, clarity, and accuracy. **Standard 5.10 is fully met.**

Standard 5.13: Transmission of individually identified test scores to authorized individuals or institutions should be done in a manner that protects the confidential nature of the scores.

Standard 5.15: When test data about a person are retained, both the test protocol and any written report should also be preserved in some form. Test users should adhere to the policies and record-keeping practice of their professional organizations.

Standard 5.16: Organizations that maintain test scores on individuals in data files or in an individual's records should develop a clear set of policy guidelines on the duration of retention of an individual's records, and on the availability and use over time, of such data.

CDE does not maintain individually identified information on students. Contractors for the development, administration, scoring, and evaluation of the CAHSEE were required to submit data confidentiality plans that were subject to legal review. Ongoing information on individual students is retained by schools and districts, subject to confidentiality restrictions in the California Education code. **Standards 5.13, 5.15, and 5.16 are fully met.**

Supporting Documentation for Tests

Section 6 of the *Standards* covers documentation requirements. Relevant standards are listed here. Nearly all of the items to be documented are discussed above. Discussion in this section is focused on documentation of these items.

Standard 6.1: Test documents (e.g., test manuals, technical manuals, user's guides, and supplemental material) should be made available to prospective test users and other qualified persons at the time a test is published or released for use.

Standard 6.2: Test documents should be complete, accurate, and clearly written so that the intended reader can readily understand the content.

Standard 6.3: The rationale for the test, recommended uses of the test, support for such uses, and information that assists in score interpretation should be documented. Where particular misuses of a test can be reasonably anticipated, cautions against such misuses should be specified.

Standard 6.4: The population for whom the test is intended and the test specifications should be documented. If applicable, the item pool and scale development procedures should be described in the relevant test manuals. If normative data are provided, the norming population should be described in terms of relevant demographic variables, and the year(s) in which the data were collected should be reported.

Standard 6.5: When statistical descriptions and analyses that provide evidence of the reliability of scores and the validity of their recommended interpretations are available, the information should be included in the test's documentation. When relevant for test interpretation, test documents ordinarily should include item level information, cut scores and configural rules, information about raw scores and derived scores, normative data, the standard errors of measurement, and a description of the procedures used to equate multiple forms.

Standard 6.9: Test documents should cite a representative set of the available studies pertaining to general and specific uses of the test.

Standard 6.14: Every test form and supporting document should carry a copyright date or publication date.

Standard 6.15: Test developers, publishers, and distributors should provide general information for test users and researchers who may be required to determine the

appropriateness of an intended test use in a specific context. … General information also should be provided for test takers and legal guardians who must provide consent prior to a test's administration.

CDE provides extensive documentation of the CAHSEE through its website and also distributes this information directly to testing coordinators in each high school district that includes grade 10. Very detailed technical documentation for each administration has been provided by the test development contractors (Yoon 2002; Smith et al., 2002; Educational Testing Service, 2003). The independent evaluator has reviewed this documentation and confirms its completeness. **All of the relevant standards in Section 6 are fully met.**

Fairness in Testing and Test Use

The second part of the *Standards* includes standards for fairness. These include general standards for fairness in testing and test use as well as standards specific to the use of tests with linguistic minorities and for individuals with disabilities who may require accommodation.

Standard 7.1: When credible research reports that test scores differ in meaning across examinee subgroups for the type of test in question, then to the extent feasible, the same forms of validity evidence collected for the examinee population as a whole should also be collected for each relevant subgroup. Subgroups may be found to differ with respect to appropriateness of test content, internal structure of test responses, the relation of test scores to other variables, or the response processes employed by individual examinees. Any such findings should receive due consideration in the interpretation and use of scores as well as in subsequent test revisions.

Standard 7.2: When credible research reports differences in the effects of construct-irrelevant variance across subgroups of test takers on performance on some part of the test, the test should be used, if at all, only for those subgroups for which evidence indicates that valid inferences can be drawn from test scores.

Standard 7.4: Test developers should strive to identify and eliminate language, symbols, words, phrases, and content that are generally regarded as offensive by members of racial, ethnic, gender, or other groups, except when judged to be necessary for adequate representation of the domain.

Validity of the CAHSEE tests is established through review of the content of the test questions. Each test question is specifically reviewed for sensitivity and fairness to different demographic groups by panels that include representatives of the relevant demographic groups. **Standards 7.1, 7.2 and 7.4 are fully met.**

Standard 7.3: When credible research reports that differential item functioning exists across age, gender, racial/ethnic, cultural, disability, and/or linguistic groups in the population of test takers in the content domain measured by the test, test developers should conduct appropriate studies when feasible. Such research should seek to detect

and eliminate aspects of test design, content, and format that might bias test scores for particular groups.

Statistical analyses to check for differential item functioning (DIF) meet industry standards. Any test question flagged for DIF is subjected to a careful review leading to a decision regarding operational use. **Standard 7.3 is fully met.**

Standard 7.5: In testing applications involving individualized interpretations of test scores other than selection, a test taker's score should not be accepted as a reflection of standing on the characteristic being assessed without consideration of alternate explanations for the test taker's performance on that test at that time.

Care has been taken to remove any irrelevant difficulties from the test form. Examinees who do not perform well at one testing session are provided several opportunities to retake the test. The State Board of Education also has established policies on accommodations, modifications, and waivers that remove barriers to students' ability to demonstrate mastery of the required standards. **Standard 7.5 is adequately met**, although additional consideration might be required if new and convincing alternate explanations for poor test performance were advanced**.**

Standard 7.7: In testing applications where the level of linguistic or reading ability is not part of the construct of interest, the linguistic or reading demands of the test should be kept to the minimum necessary for the valid assessment of the intended construct.

Some concerns about linguistic requirements for responding to mathematics questions were expressed in independent item reviews conducted by the evaluator (Wise et al, 2002). Questions are specifically reviewed for reading ability requirements prior to operational use and the test developer is continuing efforts to further simplify reading levels. **Standard 7.7 is adequately met.**

Standard 7.9: When tests or assessments are proposed for use as instruments of social, educational, or public policy, the test developers or users proposing the test should fully and accurately inform policymakers of the characteristics of the tests as well as any relevant and credible information that may be available concerning the likely consequences of test use.

The test developer provides technical documentation and regular information to CDE and SBE. An independent evaluation of test characteristics and consequences of the CAHSEE requirement is also ongoing. Regular reports are issued to the governor, legislature, State Board of Education, and the California Department of Education. **Standard 7.9 is fully met.**

Standard 7.10: When the use of a test results in outcomes that affect the life chances or educational opportunities of examinees, evidence of mean test score differences between relevant subgroups of examinees should, where feasible, be examined for subgroups for which credible research reports mean differences for similar tests. Where mean differences are found, an investigation should be undertaken to

determine that such differences are not attributable to a source of construct underrepresentation or construct-irrelevant variance. While initially the responsibility of the test developer, the test user bears responsibility for uses with groups other than those specified by the developer.

Scoring differences for relevant subgroups have been monitored and reported by CDE and by the independent evaluator. While such differences exist, every indication is that the target content is covered fully and fairly for each group. As noted above, some questions have been raised about the possible impact of reading requirements for linguistic minorities. These concerns are clearly not relevant to the section of CAHSEE that assesses reading, which all students are required to pass. Given that students have passed the reading section, there should not be problems with the reading level of mathematics. Nonetheless, the development contractor is continuing work to keep reading requirements to a minimum on the mathematics test. **Standard 7.10 is fully met.**

Standard 7.11: When a construct can be measured in different ways that are approximately equal in their degree of construct representation and freedom from construct-irrelevant variance, evidence of mean score differences across relevant subgroups of examinees should be considered in deciding which test to use.

Guidance from the original HSEE Panel and the expert experience of the test developers have led to an assessment judged to best represent the intended ELA and mathematics achievement constructs. No alternative ways of measuring these constructs have been suggested, so **Standard 7.11 is not relevant at this time.**

Standard 7.12: The testing or assessment process should be carried out so that test takers receive comparable and equitable treatment during all phases of the testing or assessment process.

As noted above, test administration procedures have been carefully standardized and training has been provided to local testing coordinators. **Standard 7.12 is fully met.**

Testing Individuals of Diverse Linguistic Backgrounds

California has a large population of students who are not native speakers of English. Consequently, section 9 of the Standards concerning testing individuals with diverse linguistic backgrounds is particularly relevant. On the other hand, by statute, part of the CAHSEE covers reading in English. This requirement limits the types of requirements that could be provided to linguistic minorities without altering the construct being assessed.

Standard 9.1: Testing practice should be designed to reduce threats to the reliability and validity of test score inferences that may arise from language differences.

As noted above, all test questions are reviewed for sensitivity and fairness for different examinee groups and for reading level requirements. Work to ensure minimal language requirements for the mathematics test is proceeding. **Standard 9.1 is adequately met.**

Standard 9.2: When credible research evidence reports that test scores differ in meaning across subgroups of linguistically diverse test takers, then to the extent feasible, test developers should collect for each linguistic subgroup studied the same form of validity evidence collected for the examinee population as a whole.

Validity evidence is based on expert judgments about content coverage. Additional judgments on language requirements and the appropriateness of the question for different groups of students also are collected. To date, there is no clear evidence that test scores have different meaning for different linguistic groups, so **Standard 9.2 is not relevant at this time.**

The remaining standards in this section cover situations where test forms in different languages are available. These standards do not apply to the CAHSEE.

Testing Individuals with Disabilities

Section 10 of the *Standards* covers requirements for testing individuals with disabilities. Relevant standards from this section are listed here. Note that the standards refer to modifications to the test and test administration procedures in a generic sense. For CAHSEE, the term modification has been reserved for changes that alter the construct being assessed. Changes that do not alter the construct are referred to as accommodations.

Standard 10.1: In testing individuals with disabilities, test developers, test administrators, and test users should take steps to ensure that the test score inferences accurately reflect the intended construct rather than any disabilities and their associated characteristics extraneous to the intent of the measurement.

Current policies and regulations covering appropriate testing accommodations for students with disabilities have clearly identified changes judged to alter the construct being measured (e.g., oral presentation of the ELA test or use of calculators on the mathematics test). These policies are consistent with policies adopted in most other states. Allowable accommodations, designed to enhance the appropriateness of scores for students with disabilities, are also consistent with common industry practice. **Standard 10.1 is fully met.**

Standard 10.2: People who make decisions about accommodations and test modification for individuals with disabilities should be knowledgeable of existing research on the effects of the disabilities in question on test performance. Those who modify tests should also have access to psychometric expertise for so doing.

Standard 10.8: Those responsible for decisions about test use with potential test takers who may need or may request specific accommodations should (a) possess the information necessary to make an appropriate selection of measures, (b) have current information regarding the availability of modified forms of the test in question, (c) inform individuals, when appropriate, about the existence of modified forms, and (d) make these forms available to test takers when appropriate and feasible.

Standard 10.10: Any test modifications adopted should be appropriate for the individual test taker, while maintaining all feasible standardized features. A test professional needs to consider reasonably available information about each test taker's experiences, characteristics, and capabilities that might impact test performance, and document the grounds for the modification.

As specified in the Individuals with Disabilities Education Act (IDEA), decisions about the appropriateness of accommodations in instruction and testing are made by the local team that works with students to develop their Individualized Education Plans (IEP). Individuals on these teams are experts on issues affecting students with disabilities. Requirements of the IEP govern the provision of accommodations on the CAHSEE. **Standards 10.2, 10.8, and 10.10 are fully met.**

Standard 10.3: Where feasible, tests that have been modified for use with individuals with disabilities should be pilot tested on individuals who have similar disabilities to investigate the appropriateness and feasibility of the modifications.

Standard 10.4: If modifications are made or recommended by test developers for test takers with specific disabilities, the modifications as well as the rationale for the modifications should be described in detail in the test manual and evidence of validity should be provided whenever available. Unless evidence of validity for a given inference has been established for individuals with the specific disabilities, test developers should issue cautionary statements in manuals or supplementary materials regarding confidence in interpretations based on such test scores.

Standard 10.5: Technical material and manuals that accompany modified tests should include a careful statement of the steps taken to modify the tests to alert users to changes that are likely to alter the validity of inferences drawn from the test score.

Standard 10.6: If a test developer recommends specific time limits for people with disabilities, empirical procedures should be used, whenever possible, to establish time limits for modified forms of timed tests rather than simply allowing test takers with disabilities a multiple of the standard time. When possible, fatigue should be investigated as a potentially important factor when time limits are extended.

Standard 10.7: When sample sizes permit, the validity of inferences made from test scores and the reliability of scores on tests administered to individuals with various disabilities should be investigated and reported by the agency or publisher that makes the modification. Such investigations should examine the effects of modifications made for people with various disabilities on resulting scores, as well as the effects of administering standard unmodified tests to them.

The preceding standards cover the development and validation of specific testing accommodations. Because of limited numbers of students in most disability categories, it is not feasible to pilot test each accommodation with separate groups of students in these categories, although students with disabilities have been included in field tests of new questions as well as in the operational administration. Separate studies of fatigue and other

factors are also not feasible, but the accommodations offered follow common practices adopted in most testing programs. All students are allowed essentially unlimited time. **Standards 10.3 through 10.7 are fully met.**

> *Standard 10.11: When there is credible evidence of score comparability across regular and modified administrations, no flag should be attached to a score. When such evidence is lacking, specific information about the nature of the modification should be provided, if permitted by law, to assist test users properly to interpret and act on test scores.*

Score reports do not indicate whether a testing accommodation was used. The reports do, appropriately, indicate if a modification that invalidates the results was used. **Standard 10.11 is fully met.**

Testing Applications: Educational Testing and Assessment

Part 3 of the *Standards* covers the standards that apply to specific types of tests. Section 13 covers educational tests and assessments. Relevant standards from this section are listed below.

> *Standard 13.1: When educational programs are mandated by school, district, state, or other authorities, the ways in which test results are intended to be used should be clearly described. It is the responsibility of those who mandate the use of tests to monitor their impact and to identify and minimize potential negative consequences. Consequences resulting from the uses of the test, both intended and unintended, should also be examined by the test user.*

In fact, the legislation establishing the CAHSEE requirement also mandated an ongoing evaluation of impact or consequences of this requirement. Regular reports on the use of test results and the impact and consequences have been prepared for relevant policy-makers, including the legislature and governor, the State Board of Education, and the Superintendent of Public Instruction. **Standard 13.1 is fully met.**

> *Standard 13.2: In educational settings, when a test is designed or used to serve multiple purposes, evidence of the test's technical quality should be provided for each purpose.*

The focus of this review is on a single use of the CAHSEE—assessing mastery of the targeted content standards. Efforts to review the technical quality of the CAHSEE as it is used for school accountability or for diagnostic purposes have been undertaken, but are not reviewed here. Thus, **Standard 13.2 is not relevant to this review.**

> *Standard 13.3: When a test is used as an indicator of achievement in an instructional domain or with respect to specified curriculum standards, evidence of the extent to which the test samples the range of knowledge and elicits the processes reflected in the target domain should be provided. Both tested and target domains should be described in sufficient detail so their relationship can be evaluated. The analyses should make*

explicit those aspects of the target domain that the test represents as well as those aspects that it fails to represent.

As noted above, test blueprints specifying the content domain were extensively debated and finally approved by the State Board. Each test form is reviewed for compliance with these blueprints. **Standard 13.3 is fully met.**

Standard 13.5: When test results substantially contribute to making decisions about student promotion or graduation, there should be evidence that the test adequately covers only the specific or generalized content and skills that students have had an opportunity to learn.

Standard 13.5 is the subject of the remainder of this report. Unfortunately, there are no clearly accepted criteria as to what constitutes adequate opportunity to learn the material on the test. Certainly, instruction covering all of the required content standards is offered in all school systems. Unfortunately, it currently appears that not all students are prepared or willing to take advantage of this instruction. **It is currently unclear whether Standard 13.5 has been fully met at this time.**

Standard 13.6: Students who must demonstrate mastery of certain skills or knowledge before being promoted or granted a diploma should have a reasonable number of opportunities to succeed on equivalent forms of the test or be provided with construct-equivalent testing alternatives of equal difficulty to demonstrate the skills or knowledge. In most circumstances, when students are provided with multiple opportunities to demonstrate mastery, the time interval between the opportunities should allow for students to have the opportunity to obtain the relevant instructional experiences.

Students have at least seven opportunities over a two and a half year period to pass the CAHSEE. **Standard 13.6 is fully met.**

Standard 13.7: In educational settings, a decision or characterization that will have major impact on a student should not be made on the basis of a single test score. Other relevant information should be taken into account if it will enhance the overall validity of the decision.

Standard 13.7 has been the subject of wide-ranging interpretations. Many argue that offering multiple opportunities to take the test satisfies the requirement of not basing an important decision on a single test score. Others argue that there are other requirements for obtaining a diploma, such as course work, that must also be met so that diplomas are not granted on test scores alone.

In fact, however, a diploma can be denied on the basis of test scores alone. Further, while providing multiple opportunities to pass the test is essential, it is also essential that there be some mechanism for consideration of other clear evidence of mastery of the required skills. The original legislation does provide for an alternate way in which students who have clearly mastered the standards can be exempted from passing the test. In addition, the policy on waivers continues to evolve. These are both ways in which other information can be

considered. Given these alternatives, we conclude that **Standard 13.7 is adequately met**. Several states with graduation test requirements have enacted additional provisions for granting waivers based on other evidence of mastery of required standards. California may wish to review such policies in considering further options for allowing students alternative ways to demonstrate the skills required for a diploma.

> *Standard 13.9: When tests scores are intended to be used as part of the process for making decisions for educational placement, promotion, or implementation of prescribed educational plans, empirical evidence documenting the relationship among particular test scores, the instructional programs, and desired student outcomes should be provided. When adequate empirical evidence is not available, users should be cautioned to weigh the test results accordingly in light of other relevant information about the student.*

As a graduation requirement, the CAHSEE is not used to place or promote students into particular programs. Consequently, **Standard 13.9 is not relevant for the intended use of the CAHSEE.**

> *Standard 13.10: Those responsible for educational testing programs should ensure that the individuals who administer and score the tests(s) are proficient in the appropriate test administration procedures and scoring procedures and that they understand the importance of adhering to the directions provided by the test developer.*

As noted above, the development contractor has produced test administration manuals, conducts test administration workshops, and monitors testing at a sample of sites. Tests are generally administered by school staff who also administer a wide range of tests and thus have considerable experience. Tests are scored by professionals trained and closely monitored by the development contractor. **Standard 13.10 is fully met.**

> *Standard 13.11: In educational settings, test users should ensure that any test preparation activities and materials provided to students will not adversely affect the validity of test score inferences.*

CDE is working to provide students and their parents and teachers with appropriate information to inform practice and instruction. Test security is maintained to prevent teaching of specific questions used in operational testing. While inappropriate test preparation will be an ongoing issue, available evidence is that **Standard 13.11 is adequately met at this time.**

> *Standard 13.12: In educational settings, those who supervise others in test selection, administration, and interpretation should have received education and training in testing necessary to ensure familiarity with the evidence for validity and reliability for tests used in the educational setting and to be prepared to articulate or to ensure that others articulate a logical explanation of the relationship among the tests used, the purposes they serve, and the interpretations of the test scores.*

CDE and the SBE have policy oversight for the development and use of the CAHSEE under requirements of the Education Code. They have contracted for test development and evaluation activities with organizations and individuals widely recognized as testing experts. These experts are charged with informing them of issues relevant to the uses of the CAHSEE. **Standard 13.12 is fully met.**

Standard 13.13: Those responsible for educational testing programs should ensure that the individuals who interpret the tests results to make decisions within the school context are qualified to do so or are assisted by and consult with persons who are so qualified.

Interpretation of test results is provided on score reports established by CDE and reviewed and approved by the SBE. Local educators are not required to provide further interpretation so **Standard 13.13 is not relevant to the use of CAHSEE as a graduation requirement.**

Standard 13.14: In educational settings, score reports should be accompanied by a clear statement of the degree of measurement error associated with each score or classification level and information on how to interpret the scores.

Measurement and, more importantly, classification error are both described extensively in technical documentation provided by the test developers and by the evaluators. **Standard 13.14 is adequately met.**

Standard 13.15: In educational settings, reports of group differences in test scores should be accompanied by relevant contextual information, where possible, to enable meaningful interpretation of these differences. Where appropriate contextual information is not available, users should be cautioned against misinterpretation.

The primary use of CAHSEE results is for decisions about individual students. Results for different demographic groups are presented annually and evaluation reports have investigated some factors behind such differences. **Standard 13.15 is adequately met.**

Standard 13.16: In educational settings, whenever a test score is reported, the date of test administration should be reported. This information and the age of any norms used for interpretation should be considered by test users in making inferences.

All score reports clearly indicate the test administration data. Normative information is not relevant to the use of CAHSEE results as a graduation requirement. **Standard 13.16 is fully met.**

Standard 13.17: When change or gain scores are used, such scores should be defined and their technical qualities should be reported.

Gain scores are not relevant to the use of CAHSEE as a graduation requirement so **Standard 13.17 is not applicable to this use.** As noted above, however, gain scores may be used by students, teachers, and parents in making decisions about appropriate remediation

strategies. Further investigation of the technical qualities of such gain scores may be warranted.

> *Standard 13.19: In educational settings, when average or summary scores for groups of students are reported, they should be supplemented with additional information about the sample size and shape or dispersion of score distributions.*

The primary reports for groups of students are in terms of the percent passing each section of the CAHSEE. The percent passing includes complete information on the underlying dichotomous distribution, so **Standard 13.19 is adequately met.**

Summary

Each of the standards for test development and use adopted by a joint committee of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education was reviewed. *For use as a high school graduation requirement, the CAHSEE meets all of the relevant standards, with the possible exception of Standard 13.5*. The one exception concerns whether students have had adequate opportunity to learn the material covered by the CAHSEE. Information on this issue is the focus of the remainder of this report.

One other particular standard, 13.7, requiring that important decisions not be made based on a single test score, is open to some interpretation. Expanded options for ways students might meet the CAHSEE requirement could further strengthen California's case for compliance with this standard.

The focus of the current investigation was on whether the CAHSEE meets standards for use as a high school graduation requirement. There are, of course, other possible or contemplated uses of the CAHSEE score information. These include use of the CAHSEE in the state's academic performance index (API) used for school accountability, use of the CAHSEE scores together with additional performance level standards to satisfy requirements of the No Child Left Behind legislation, and diagnostic interpretation of subscores and score gains. Further review and documentation would likely be required to conclude that these uses of the CAHSEE are in full compliance with the *Standards*. We've noted specific issues with some of these uses in the text above.